RESEARCH STATEMENT

Kai Zhang (twinsen@cse.ust.hk)
Department of Computer Science & Engineering
Hong Kong University of Science and Technology

My research work is centered around *clustering*, the most basic form of unsupervised learning. It extends to several closely related areas including: *probabilistic mixture models*, *manifold leaning and dimension reduction*, *low rank approximation*, *maximum margin clustering* and *semi-supervised learning*. An important theme of my work is to scale up the currently existing, but computationally demanding algorithms to large scale, real world problems. This can benefit a variety of applications in data mining, computer vision and information retrieval.

Probabilistic Mixture Models

Mixture model is among the most popular and well studied probabilistic models, which is widely used in discriminant analysis, hierarchical clustering, density estimation, and applications such as image segmentation, gene-expression analysis, and speech processing. Two core problems related are

1. *Learning: how to learn the model parameters for data modelling?*

2. *Model simplification: how to obtain a simple mixture model?*

My current work is on simplifying mixture models. Many learning algorithms involve mixture models in the testing (or learning) phase, such as Parzen classifier, SVM testing, Bayesian filtering, and switching linear dynamic models. A simpler model allows much faster training/testing. Also, model simplification can be interpreted as component clustering, and this *functional clustering* frame can solve interesting practical problems such as image categorization [1], sensor network clustering [2], and statistical software debugging [2]. In [10, 14], I proposed a model simplification scheme by minimizing the upper bound of the L_2 approximation error between the original and the simplified model, through iterative component grouping and local function approximation. The algorithm can simultaneously solve the problem of *model simplification* and *component clustering*, and is more robust and accurate than existing algorithms [1, 2] in nonparametric density estimation, mean shift image segmentation [4], and SVM testing.

Undergoing project. By applying the distance minimization idea [10] to the approximation of the underlying true distrbution f, a new non-parametric data modelling scheme can be derived. Unlike traditional Mixture Model learning algorithms such as Expectation Maximization, this distance-minimization framework can alleviate the difficulty of model selection by sequentially inserting base kernels. On the other hand, by controlling the smoothness of base kernels, the risk of over-fitting can be prevented. It also demonstrates interesting connection with the pre-image problem in kernel methods.

Manifold Learning and Dimension Reduction

Exploring intrinsic structures of real world data (such as customer record, blogs, internet, bioinformatics data, etc.) and embedding them into their intrinsic, possibly low-dimensional space is of increasing interest to data analysis and visualization, and remains an active domain in data mining applications. A key numerical procedure needed here is the eigenvalue decomposition which is often computationally expensive. The Nyström method [3, 5] is a sampling based technique widely used in approximating the large kernel eigen-systems. In [11, 13], I extend the Nyström method a more general, density weighted formulation. The basic step is to approximate the $n \times n$ kernel matrix with a block-wise constant version, whose eigen-system can be obtained easily in $O(m^3)$ time ($m \ll n$). The eigen-system is then refined through the density weighted Nyström Extrapolation. Empirically, the weighted Nyström method is more accurate by orders of magnitudes than the standard Nyström method, and demonstrates encouraging performance in kernel principal component analysis, normalized cut, and image segmentation [11].

Low Rank Approximation & Matrix Factorization

Ongoing project. Low rank approximation [5] is an important technique in alleviating the memory and computational burdens of kernel method that is playing a key role in machine learning. By examining the matrix completion view of the density-weighted Nyström method [11, 13], an improved Nyström low rank approximation scheme can be derived, with a deterministic error analysis of approximation. This work has been submitted and is currently under review. Empirically, it demonstrate supervisor performance than a number of state-of-the-art techniques such as incomplete Choleskey decomposition, the Nyström method [5], and some advanced probabilistic sampling approaches [4].

Ongoing project. Eigenvalue decomposition is a special form of the more general matrix factorization technique, Singular Value Decomposition (SVD). Its importance and wide range of applications need not to be further emphasized here. Currently, based on the quantization of systems of linear equations, I have designed an efficient approximate SVD solver. In comparison with the randomized algorithms that focus on various probabilistic sampling schemes, such as the representative work in [4, 8], our approach again demonstrates superior numerical behavior.

Maximum Margin Clustering

Maximum Margin Clustering reflects a recent interest in development of clustering algorithms, i.e., clustering can be simultaneously achieved with maximizing the separation margin between data clusters. Unfortunately, existing works are all based on the Semi-definite Programming [6, 7], and the scale of the data set that can be handled so far is extremely limited. In [9, 15], I proposed an efficient approach that performs alternating optimization directly on the original non-convex problem, which avoids the SDP relaxations. We use the symmetric Laplacian loss to replace the hinge loss in the SVM in the inner optimization, which greatly alleviates the premature convergence. Empirically, our approach is several

orders of magnitude times faster than existing SDP approaches with very competitive performance [9].

Semi-supervised Learning

Ongoing project. Semi-supervised learning is an important learning paradigm that greatly alleviates the cost of supervised learning in terms of reducing the need for expensive, labelled data. Not surprising, though, existing semi-supervised learning techniques can be quite expensive due to the lack of an efficient mechanism in summarizing/utilizing the valuable information carried among the large number of unlabelled patterns. By using the low rank approximation scheme that is extended from [11], I have designed an efficient semi-supervised inference algorithm called *Prototype Vector Machine* (PVM), which utilizes the sparse low-rank representation of the kernel matrix to reduce the number of variables in semi-supervised inference. Preliminary results on semi-supervised clustering and interactive image segmentation have shown encouraging performance.

References

[1] Jacob Goldberger and Sam Roweis. Hierarchical Clustering of a Mixture Model. In the *Neural Information Processing Systems 18*, 2005.

[2] Jason V. Davis and Inderjit Dhillon. Differential Entropic Clustering of Multivariate Gaussians. In the *Neural Information Processing Systems 19*, 2006.

[3] Charless Fowlkes, Serge Belongie, Fan Chung and Jitendra Malik. Spectral Grouping using the Nyström Method. In *IEEE Transactions on Pattern Analysis and Machine*, 26 (2), pp. 214-225.

[4] Petros Drineas and Michael W. Mahoney. On the Nyström Method for Approximating a Gram Matrix for Improved Kernel-Based Learning. In *Journal of Machine Learning Research* 6, 2153-2175, 2005.

[5] Christopher K. I. Williams and Matthias Seeger. Using the Nyström Method to Speed Up Kernel Machines. In the *Advances in Neural Information Processing Systems 13*, 2000.

[6] Lin L. Xu, James Neufeld, Bryce Larson, Dale and Schuurmans. Maximum Margin Clustering. In the *Advances in Neural Information Processing Systems 16*, 2004.

[7] Hamed Valizadegan and Rong Jin. Generalized Maximum Margin Clustering and Unsupervised Kernel Learning. In the *Advances in Neural Information Processing Systems 19*, 2006.

[8] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo Algorithms for Matrices II: Computing a Low Rank Approximation to a Matrix. In the *SIAM Journal of Computing*, 2005.

[9] Kai Zhang, Ivor W. Tsang, James T. Kwok. Maximum Margin Clustering Made Practical. In the *24th International Conference on Machine Learning* , 2007.

[10] Kai Zhang, James T. Kwok. Simplifying Mixture Models Through Function Approximation. In the *Neural Information Processing Systems 19*, 2006.

[11] Kai Zhang, James T. Kwok. Block-Quantized Kernel Matrix for Fast Spectral Embedding. In the *23rd International Conference on Machine Learning*, 2006.

[12] Kai Zhang, M. Tang, J. T. Kwok. Applying Neighborhood Consistency for Fast Clustering and Kernel Density Estimation. In the *International Conference on Computer Vision and Pattern Recognition*, 2005.

[13] Kai Zhang, James T. Kwok. Density-Weighted Nyström Method for Computing Large Kernel Eigen-Systems, submitted to *Neural Computation*.

[14] Kai Zhang, James T. Kwok. Simplifying Mixture Models Through Function Approximation, submitted to *Journal of Machine Learning Research*.

[15] Kai Zhang, Ivor W. Tsang, James T. Kwok. Maximum Margin Clustering Made Practical, submitted to *IEEE Transactions on Neural Networks*.